PadhAI: Backpropagation - the full version

One Fourth Labs

Computing derivatives w.r.t Hidden Layers
Part 3

1. Consider the next layer a_i
   a. $\frac{\partial L(\theta)}{\partial a_{ij}} = \frac{\partial L(\theta)}{\partial h_{ij}} \frac{\partial h_{ij}}{\partial a_{ij}}$
   b. The first derivative is what we computed in part 2
   c. We need to compute the second derivative $\frac{\partial h_{ij}}{\partial a_{ij}}$
   d. We know that h_{ij} is simply the application of an activation function (sigmoid, tanh etc) to a_{ij}
   e. So it can be rewritten as $\frac{\partial h_{ij}}{\partial a_{ij}} = g'(a_{ij})$ where $h_{ij} = g(a_{ij})$ and $g'(a_{ij})$ is its derivative
2. $\frac{\partial L(\theta)}{\partial a_{ij}} = \frac{\partial L(\theta)}{\partial h_{ij}} g'(a_{ij})$
3. The full gradient can be written as
   a.

$$\nabla_{a_i} L(\theta) = \begin{bmatrix} \frac{\partial L(\theta)}{\partial h_{i1}} g'(a_{i1}) \\ \cdot \\ \cdot \\ \cdot \\ \frac{\partial L(\theta)}{\partial h_{in}} g'(a_{in}) \end{bmatrix}$$

   b. This vector is the element-wise product of two vectors $\nabla_{h_i} L(\theta)$ and $[\ldots, g'(a_{ik}), \ldots]$ (which is a vector of derivations of the activation function w.r.t the pre-activation layer. They are both vectors of n-terms
4. Thus $\nabla_{a_i} L(\theta) = \nabla_{h_i} L(\theta) \odot [\ldots, g'(a_{ik}), \ldots]$ ($\odot$ refers to element-wise multiplication)
5. This formula can be applied to any of the hidden layers